UNITED STATES DISTRICT COURT
EASTERN DISTRICT OF NEW YORK
----------X
UNITED STATES OF AMERICA,

MEMORANDUM AND ORDER

Plaintiff,

07-cv-2067 (NGG) (RLM)

-and-

THE VULCAN SOCIETY, INC., MARCUS HAYWOOD, CANDIDO NUÑEZ, and ROGER GREGG,

Plaintiff-Intervenors,

-against-

THE CITY OF NEW YORK,

Defendant.
----------X

NICHOLAS G. GARAUFIS, United States District Judge.

From 1999 to 2007, the New York City Fire Department used written examinations that had discriminatory effects on minority applicants and failed to test for relevant job skills. In July 2009 and January 2010, this court held that New York City's use of these examinations discriminated against black and Hispanic applicants in violation of Title VII of the Civil Rights Act of 1974, and against black applicants in violation of the United States Constitution. Today, the court concludes that the Fire Department's current written examination, Exam 6019, does not comply with Title VII. As a result, the court temporarily enjoins the City from using Exam 6019 to appoint entry-level firefighters.

The court's analysis is complex, but its conclusion is simple. The City has not shown that the current examination identifies candidates who will be successful firefighters. Because the test questions do not measure the abilities required for the job of entry-level firefighter, the

examination cannot distinguish between qualified and unqualified candidates, or even between more and less qualified candidates. In the words of the Second Circuit Court of Appeals, the examination "satisfies a felt need for objectivity, but it does not necessarily select better job performers."[1] What the examination does do is screen and rank applicants in a manner that disproportionately excludes black and Hispanic applicants. As a result, hundreds of minority applicants are being denied the opportunity to serve as New York firefighters, for no legitimate or justifiable reason.

The court is not alone in its opinion that Exam 6019 fails to test for useful skills and abilities. The firefighters and fire lieutenants who reviewed the examination before it was administered overwhelmingly agreed that large portions of the exam should not be used. They also offered the following comments:

> I feel all these questions are unfair. They have nothing to do with an entry-level exam.
>
> No good. These questions should be used to help in a psychological profile of the applicant. They should not be used for an entrance exam.
>
> This should not be part of the test. It is subjective.
>
> Prior firehouse knowledge needed. Members/candidates with prior firehouse or fire ground knowledge will have a great unfair advantage compared to the general public.[2]

The City ignored the opinions of its own firefighters when creating Exam 6019. As a result, the City administered an invalid, discriminatory exam to nearly 22,000 job applicants.

This court previously ordered the parties to begin constructing a new, valid firefighter selection procedure under the guidance of Special Master Mary Jo White. The question now is whether and how the City may use Exam 6019 to appoint new firefighters in the interim. At this

[1] Guardians Assoc. of New York City Police Dept., Inc. v. Civil Service Comm'n., 630 F.2d 79, 100 (2d Cir. 1980).

[2] See Pl. Ex. 20.

time, the court does not have enough information to decide on a permanent course of action. The City asserts, without offering any documentary or testimonial support, that it needs to hire a new firefighter class immediately. But the City cancelled its last class of appointees in 2009, and earlier this year Mayor Michael Bloomberg advocated closing 20 fire companies and reducing staffing in 60 additional engine companies.[3] Before this court can permit the City to use Exam 6019 in any manner, the City must explain what has changed and why the need to appoint a few hundred rookie firefighters using and invalid test outweighs the need to avoid racial discrimination in municipal hiring. Accordingly, this court will hold a hearing as soon as possible to consider the remedial measures it should take in light of today's decision and the City's needs.

I. BACKGROUND

A. Litigation History

Between 1999 and 2008, the City used two competitive examination processes, Exam 7029 and Exam 2043, to screen and select applicants for entry-level firefighter positions. In 2002 and 2005, the Intervenors filed charges with the Equal Employment Opportunity Commission ("EEOC"), alleging that the exams violated Title VII. (See Int. D.I. 56.1 ¶¶ 4-5.) In 2004, the EEOC determined that Exam 7029 adversely affected black applicants and was invalid. In 2005, the EEOC made the same determination regarding Exam 2043. (Id.) The City refused to conciliate, and the EEOC referred the charges to the United States Department of Justice ("DOJ"). (Id. ¶ 6 & Ex. K.) The DOJ filed the instant lawsuit in May 2007. (See Compl.(Docket Entry # 1).)

[3] See Docket Entry # 502, Ex. A.

The City began developing its current test, Exam 6019, in August 2006, after the EEOC determined that Exams 7029 and 2043 were invalid. (See 6019 Test Development Report (Def. Ex. A-3) ("Test Dev. Rep.") 2.) The City administered Exam 6019 on January 20, 2007. (Pl. Ex. 1.) Approximately 21,983 candidates completed the exam, and 21,235 candidates passed. (Exam 6019 Analyses and Scoring Report (Def. Ex. A-5)("Exam Analysis") 3; Pl. Ex. 4a.) The City established the Exam 6019 "eligibility list" – i.e., the rank-order list of those who passed – in June 2008, and hired its first (and to date, only) academy class off the list in July 2008. (Seeley Decl. (Docket Entry # 316-1), Ex. C; HT 228-29.)

In July 2009, this court held that the City's use of Exams 7029 and 2043 as pass/fail and rank-ordering devices constituted disparate-impact discrimination in violation of Title VII. See United States v. City of New York, 637 F. Supp. 2d 77 (E.D.N.Y. 2009) ("Disparate Impact Opinion" or "D.I. Op."). In January 2010, this court also held that the City's actions constituted intentional discrimination in violation of Title VII and the Fourteenth Amendment. United States v. City of New York, 683 F. Supp. 2d 225 (E.D.N.Y. 2010). Following these decisions, the court issued a preliminary relief order directing the parties to take certain actions to begin remedying the City's violations. See United States v. City of New York, 681 F. Supp. 2d 274 (E.D.N.Y. 2010) ("Initial Remedy Order"). Among other things, the court directed the parties to prepare for a hearing (the "6019 Hearing") regarding the validity of Exam 6019, which in turn would determine whether and how the City could hire from the Exam 6019 eligibility list on an interim basis while a new, valid selection procedure was being developed. Id. at 278.

Under the supervision of Magistrate Judge Roanne Mann and Special Master White,[4] the parties engaged in a lengthy and occasionally contentious discovery process in preparation for

[4] The court thanks Special Master White for her invaluable pro bono assistance in this matter.

the 6019 Hearing. On June 29, 2010, the City informed the court that it intended to hire approximately 300 firefighters from the Exam 6019 eligibility list, to initiate the new academy class in either the last week of August 2010 or the first week of September 2010, and to begin notifying successful candidates approximately 30 days in advance of the start of the class. (Docket Entry # 456.) Accordingly, the court and the Special Master accelerated the schedule for the 6019 Hearing. The parties submitted voluminous pre-hearing briefing and numerous exhibits. Over the course of a two-day hearing on July 20 and 21, 2010, the court heard testimony from the City's test-construction expert, Dr. Catherine Cline; the Plaintiffs' test-construction expert, Dr. David P. Jones; and Donay Queenan, the FDNY's Assistant Commissioner for Human Resources. This opinion now follows.

B. The Exam 6019 Hiring Process

Exam 6019 is an objectively scored paper-and-pencil test consisting of 195 multiple-choice questions. (Pl. PFF ¶ 3.) The exam comprises three components: a timed component, a "situational judgment exercise" ("SJE") component, and a cognitive component. (Id.) The City weighted the scores on these components differently, with the cognitive component weighted the most and the timed component weighted the least. (Pl. Ex. 2.) The City scaled each candidate's combined score so that candidates who correctly answered all questions received a score of 100, and candidates who correctly answered 70% of all questions received a score of 70. (Pl. PFF ¶ 9; Exam Analysis 5.)

The City used a cutoff score of 70 to determine which candidates passed Exam 6019. (Pl. Ex. 1.) Candidates who failed the exam were excluded from further consideration for the job. The City calculated each passing candidate's "Adjusted Final Average" by adding any applicable

residency, veteran, and legacy bonus points to the candidate's exam score.[5] (PL. PFF ¶ 15.) The City then assigned each candidate a list number (or rank) based on the candidate's Adjusted Final Average, with the lowest list numbers (i.e., the highest ranks) assigned to the candidates with the highest Adjusted Final Averages. Candidates with the same Adjusted Final Average were ranked based upon their Social Security numbers. (Id. ¶ 17.)

Candidates' exam scores and resulting list ranking determine the order in which they are processed for hiring. Candidates are invited to take the Candidate Physical Ability Test ("CPAT") based on their rank on the Exam 6019 eligibility list. (Id. ¶ 21.) To be appointed, candidates passing the CPAT also have to appear on a certification list, meet all requirements for appointment set forth in the Exam 6019 notice of examination, and pass a medical and psychological examination. (Id.) Because the City hires firefighters in classes – typically between 150 and 300 hires at a time – it processes candidates off of the eligibility list in large groups, as many as 1,000 at a time. (6019 Hearing Transcript ("HT") 227-30.) The candidates who are found to be qualified are hired in rank-order off of the eligibility list, meaning that a candidate who has completed all steps in the selection process may still not be hired if the City fills its academy class before the candidate's list number is reached. (Pl. PFF ¶ 26.)

Since establishing the Exam 6019 eligibility list in June 2008, the City has hired only one academy class from it. The lowest-ranked applicant actually appointed by the City thus far was ranked 834th on the eligibility list. (Siskin Decl. (Docket Entry # 316-3) ¶ 11 & n.5.)

II. THE COURT'S REMEDIAL AUTHORITY

Congress enacted Title VII "to assure equality of employment opportunities and to eliminate those discriminatory practices and devices which have fostered racially stratified job

[5] In contrast to the scoring process for Exams 7029 and 2043, the City did not consider any measure of candidates' physical abilities when calculating the Adjusted Final Average on Exam 6019. (Pl. PFF ¶ 18.)

environments to the disadvantage of minority citizens." McDonnell Douglas Corp. v. Green, 411 U.S. 792, 800 (1973). In order to meet this sweeping mandate, Congress "took care to arm the courts with full equitable powers." Albemarle Paper Co. v. Moody, 422 U.S. 405, 418 (1975). Title VII authorizes district courts to choose from a wide spectrum of remedies for illegal discrimination, ranging from compensatory relief such as back pay to "affirmative relief" such as the imposition of hiring quotas. See 42 U.S.C. 2000e-5(g); Berkman v. City of New York, 705 F.2d 584, 595-96 (2d Cir. 1983); see also Rios v. Enterprise Asso. Steamfitters Local 638 etc., 501 F.2d 622, 629 (2d Cir. 1974) ("Once a violation of Title VII is established, the district court possesses broad power as a court of equity to remedy the vestiges of past discriminatory practices.").

While certain types of relief authorized by Title VII are discretionary and context-specific, other types are essentially requisite. Compare United States v. Starrett City Associates, 840 F.2d 1096, 1102 (2d Cir. 1988) (racial quotas appropriate only in limited circumstances) with Albemarle Paper Co., 422 U.S.at 422 (back pay should be awarded in almost all circumstances). In particular, once liability for racial discrimination has been established, the district court "has not merely the power but the duty" to "bar like discrimination in the future." Id. at 418 (quoting Louisiana v. United States, 380 U.S. 145, 154 (1965)). This so-called "compliance relief" is designed to assure future compliance with Title VII. Berkman, 705 F.2d at 595. In the context of discriminatory testing regimes, such relief involves "restricting the use of an invalid exam, specifying procedures and standards for a new valid selection procedure, and authorizing interim hiring that does not have a disparate racial impact." Guardians Assoc. of New York City Police Dept., Inc. v. Civil Service Comm'n, 630 F.2d 79, 108 (2d Cir. 1980) ("Guardians"); see also id. at 109 ("Once an exam has been adjudicated to be in violation of Title

VII, it is a reasonable remedy to require that any subsequent exam or other selection device receive court approval prior to use."). In this case, the City has stopped using Exams 2043 and 7029, and the court has already ordered the parties to begin constructing a new, valid selection procedure for entry-level firefighters, see February 24, 2010 Docket Entry. The court's remaining duty is to determine whether the City's interim hiring based on Exam 6019 is compatible with Title VII.

Before undertaking this analysis, the court offers a word about its methodology. The existence of a Title VII violation affords the court broad equitable remedial authority, and Guardians suggests that courts are free to dismantle interim hiring procedures based solely on their disparate impact. Nonetheless, this court believes that the appropriate course is to subject Exam 6019 to a standard Title VII analysis – that is, to assess both the exam's racial impact and whether it is job-related or consistent with business necessity – before ruling on its use. In part, this belief stems from the simple intuition that the best way to police the City's compliance with Title VII is to actually measure the City's actions against the statute's requirements. Another reason, however, is that the law in this area is in a state of flux. In Ricci v. DeStefano, 129 S. Ct. 2658 (2009), the Supreme Court held that an employer may not, consistent with the disparate-treatment provisions of Title VII, set aside the results of an exam that disparately impacts employee candidates without a strong basis in evidence that the exam actually violates Title VII. Id. at 2673-77. Ricci's specific holding does not control here, since a federal court attempting to remedy identified discrimination enjoys far more authority than an employer attempting to remedy potential discrimination. See, e.g., Albemarle Paper Co., 422 U.S. at 418. Nonetheless, Ricci announces general principles that could as easily apply to courts as employers, and this court is hesitant to ignore them absent further guidance from the Supreme Court or the Second

Circuit. One of these principles is that government actors – who are bound by the Equal Protection Clause, which is analogous to Title VII's disparate-treatment provision – must identify, rather than extrapolate, the existence of a Title VII violation before taking race-conscious remedial action such as setting aside or restricting the use of exam results. As stated before, a federal court operating in a remedial setting may well be exempt from this principle, but the court sees no reason to use this as a test case. Therefore, the prudent step is to evaluate Exam 6019 using a Title VII analysis.

At the same time, the court does not wish to overstate its actions or overstep its authority. Exam 6019 is not the subject of this lawsuit, and no party has filed an independent action alleging that it is unlawful. This court is only reaching Exam 6019 as an exercise of its remedial jurisdiction, and while that jurisdiction is broad, it is not unlimited. Thus, as this court observed previously,

> [It is not] necessary (or even appropriate) for this court to conclusively determine the legality of Exam 6019 at this juncture. The court simply wants to know, before it decides on an interim hiring remedy, whether the pass/fail and rank-ordering uses of Exam 6019 have a disparate impact on black and Hispanic test-takers, and if so, whether those uses are job-related. The exam's precise legal status, its superiority or inferiority to alternate procedures, and the rights of the candidates who took it are all important questions that may be taken up elsewhere. But they are ancillary to the question facing the court, which is what use, if any, should be made of the Exam 6019 eligibility list in the event that the FDNY begins hiring firefighters before a new examination is developed.

United States v. City of New York, 2010 U.S. Dist. LEXIS 31397, at *5-6 (E.D.N.Y. Mar. 31, 2010). Therefore, although this court is analyzing Exam 6019 under a Title VII liability framework, the freestanding "legality" of Exam 6019 is not at issue, and the court's conclusions are meant only to guide its future remedial actions.

III. THE PLAINTIFFS' PRIMA FACIE CASE

Courts assess disparate-impact discrimination under Title VII using a three-step burden-shifting framework. A plaintiff must first "establish by a preponderance of the evidence that the employer 'uses a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex, or national origin.'" Robinson v. Metro-North Commuter R.R. Co., 267 F.3d 147, 160 (2d Cir. 2001) (quoting 42 U.S.C. § 2000e-2(k)(1)(A)(i)). "To make this showing, a plaintiff must (1) identify a policy or practice, (2) demonstrate that a disparity exists, and (3) establish a causal relationship between the two." Id. at 160. If the plaintiff succeeds, the burden shifts to the defendant to demonstrate that the challenged practice or policy is "job related for the position in question and consistent with business necessity." 42 U.S.C. § 2000e-2(k)(1)(A)(i). The burden then shifts back to the plaintiff "to establish the availability of an alternative policy or practice that would also satisfy the asserted business necessity, but would do so without producing the disparate effect."[6] Robinson, 267 F.3d at 161 (citing 42 U.S.C. § 2000e-2(k)(1)(A)(ii), (C)).

Statistics alone can make out a prima facie case, provided they "reveal[] a disparity substantial enough to raise an inference of causation." EEOC v. Joint Apprenticeship Comm. of Joint Indus. Bd. of Elec. Indus., 186 F.3d 110, 117 (2d Cir. 1999). The plaintiffs' statistical evidence "must reflect a disparity so great that it cannot be accounted for by chance." Id.; see also Robinson, 267 F.3d at 160. In this case, Plaintiffs have demonstrated – and the City concedes – that the City's use of Exam 6019 as a pass/fail and rank-ordering device results in a statistically significant racial disparity sufficient to make out a prima facie case of disparate impact discrimination. See Initial Remedy Order, 681 F. Supp. 2d at 295 & n.2; see also HT 6-7.

[6] In order to expedite discovery, the Plaintiffs agreed that they would not attempt to make a showing on this step during the 6019 Hearing (see Docket Entry ## 411-12), and the court ruled that it would deem the issue waived when deciding whether and how the City could use Exam 6019, see United States v. City of New York, 2010 U.S. Dist. LEXIS 31397, at *6 & n.1 (E.D.N.Y. Mar. 31, 2010).

Plaintiffs' expert, Dr. Bernard Siskin, has demonstrated that the City's pass/fail use of Exam 6019 with a cutoff score of 70 has a statistically significant adverse impact upon black and Hispanic applicants. The disparity between the pass rates of black applicants and the pass rate of white applicants is equivalent to 22.70 units of standard deviation.[7] (Siskin Decl. ¶ 7.) The practical effect of this disparity is that 298 black candidates who would not have failed the examination but for the disparity were eliminated from consideration for the position of entry-level firefighter. (Id. ¶ 8.) The disparity between the pass rates of Hispanic applicants and the pass rate of white applicants is equivalent to 11.35 units of standard deviation. (Id. ¶ 9.) As a practical matter, this disparity eliminated 132 Hispanic candidates from consideration. (Id. ¶ 8.)

Dr. Siskin's analysis also demonstrates that the City's use of Exam 6019 as a rank-ordering device has a statistically significant adverse impact on black and Hispanic applicants, and that this adverse impact will only worsen as the City continues to use the eligibility list to hire firefighters. As Dr. Siskin points out, many applicants who nominally passed Exam 6019 have effectively failed the exam because they did not score high enough to actually be hired. Currently, the lowest-ranked applicant who has been appointed as a firefighter was ranked 834th on the Exam 6019 eligibility list.[8] (Id. ¶ 11 & n.5.) Based on the number of firefighters that the City has hired off the 6019 list to date, Dr. Siskin calculates that 4.10% of white test-takers and 2.36% of black test-takers scored high enough to actually be appointed. The disparity in the effective pass rates of white and black applicants is equivalent to 5.01 units of standard

[7] A full description of the use and probative value of statistical-significance and standard-deviation analysis in employment discrimination cases can be found in the Disparate Impact Opinion. See D.I. Op., 637 F. Supp. 2d at 87, 94. As a general matter, standard-deviation analysis is used to determine the likelihood that the difference between observed and expected values could occur randomly. The larger the standard deviation, the less likely it is that a result is the product of chance.

[8] Because there are gaps in the list numbers on the Exam 6019 eligibility list, the 834th highest-ranked applicant had a list number of 886. (Siskin Decl. ¶ 11.)

deviation. Dr. Siskin estimates that, absent the disparity, 21 additional black applicants would have been appointed from the Exam 6019 eligible list. (Id. ¶ 11.) If the City continues to use the Exam 6019 eligible list in rank-order and reaches twice as far down the list as it currently has, the disparity between the effective pass rates of white and black applicants will be equivalent to 5.94 units of standard deviation, representing an estimated 35 additional black applicants who would be hired absent the disparity. (Id. ¶ 12.) If the City reaches three times as far down the list, the disparity between the effective pass rates of white and black applicants will be equivalent to 8.11 units of standard deviation, representing an estimated 58 additional black applicants who would be hired absent the disparity. (Id. ¶ 13.) And if the City reaches four times as far down the list, the disparity between the effective pass rates of white and black applicants will be equivalent to 9.59 units of standard deviation, representing an estimated 74 additional black applicants who would be hired absent the disparity.[9] (Id. ¶ 13.)

These statistics are sufficient to establish a prima facie case of disparate-impact discrimination. As noted previously in this case, "[t]he Second Circuit has repeatedly recognized that standard deviations of more than 2 or 3 units can give rise to a prima facie case of disparate impact because of the low likelihood that such disparities have resulted from chance." D.I. Op., 637 F. Supp. 2d at 93 (citing Malave v. Potter, 320 F.3d 321, 327 (2d Cir. 2003); Waisome v. Port Auth. of N.Y. & N.J., 948 F.2d 1370, 1376 (2d Cir. 1991); Ottaviani v. State University of New York, 875 F.2d 365, 372 (2d Cir. 1989); Guardians, 630 F.2d at 86); see also Hazelwood School Dist. v. United States, 433 U.S. 299, 309 n.14 (1977) ("general rule" in employment discrimination cases with large samples is that "if the difference between the expected value and

[9] According to Dr. Siskin, "[i]f the City continues to use the Exam 6019 eligibility list in rank order, and reaches three times as far down the list as it [has], the disparity between the effective pass rate of white and Hispanic applicants also becomes statistically significant and remains statistically significant at four and five times as far down the list as the City [has] reached . . . ." (Siskin Decl. n.6.)

the observed number is greater than two or three standard deviations, then the hypothesis that [employees] were hired without regard to race would be suspect"). The calculated standard deviations in this case range from 5.01 to 22.7 units, well in excess of the Second Circuit's benchmark. (Siskin Decl. ¶¶ 7, 11.)

The City does not contest Dr. Siskin's results or methodology, and conceded at the 6019 Hearing that the Plaintiffs' evidence creates a prima facie case of disparate-impact discrimination. (See HT 6-7.) Accordingly, the dispositive question is whether the City has carried its burden of demonstrating that Exam 6019 is job-related and justified by business necessity.

IV. THE CITY'S BUSINESS NECESSITY DEFENSE

Because Plaintiffs have shown that the City's uses of Exam 6019 disparately impacts black and Hispanic candidates, the City bears the burden of showing that those uses are justified by legitimate business and job-related considerations.[10] See Gulino v. New York State Educ. Dep't, 460 F.3d 361, 385 (2d Cir. 2006). In Gulino, the Second Circuit explained the business necessity defense as follows:

> [T]he basic rule has always been that "discriminatory tests are impermissible unless shown, by professionally acceptable methods, to be predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated." This rule operates as both a limitation and a license for employers: employers have been given explicit permission to use job related tests that have a disparate impact, but those tests must be "demonstrably a reasonable measure of job performance."

[10] "Though the terms 'business necessity' and 'job related' appear to have semantic differences, they have been used interchangeably by the courts." Gulino v. New York State Educ. Dep't, 460 F.3d 361, 382 (2d Cir. 2006).

460 F.3d at 383 (quoting Albemarle Paper Co. v. Moody, 422 U.S. 405, 431, 426 (1975)). In EEOC parlance, an exam that accurately predicts on-the-job performance is said to be "valid," while an exam that does not is invalid. See, e.g., 29 CFR § 1607.5.

The validity of an exam turns in large part on the process by which it was constructed. See Guardians, 630 F.2d at 95. Accordingly, the court begins by reviewing Exam 6019's construction before turning to the substantive legal standards that govern whether the exam is job-related under Title VII.

A. Creation of Exam 6019

Dr. Catherine Cline, Ph.D., a per diem employee in the City's Department of Citywide Administrative Services ("DCAS"), was primarily responsible for developing Exam 6019. (HT 62-63.) Dr. Cline also served as the City's expert witness at the 6019 Hearing. One of the City's liability-phase experts, Dr. Philip Bobko, Ph.D., assisted Dr. Cline in developing the exam. (Id. 63.)

1. Deriving a List of Tasks and Abilities

Dr. Cline began by compiling a list of tasks that firefighters commonly perform and a list of abilities that firefighters should possess. In May 2006, Dr. Cline convened two panels consisting of six to ten firefighters and supervisors. (Job Analysis Report (Def. Ex. A-2) at DCAS-08652.) The panelists were asked to review the existing task list, which had been left over from a prior exam, and an ability list that Drs. Cline and Bobko had created based on similar lists created by the United States Department of Labor and experts in the field. (Id. at DCAS-08652-53; HT 36-37.) Over the course of the review sessions, the panelists altered the wording and ordering of some tasks and generated examples of how the listed abilities might be used on the job. (Id. at DCAS-08653-54; HT 38-39.)

After incorporating the review panelists' suggestions, Dr. Cline used the revised task and abilities lists and the abilities examples to draft a Job Analysis Questionnaire (JAQ). (Job Analysis Report at DCAS-08655.) In June 2006, DCAS administered the JAQ to 343 randomly selected firefighters. (Id. at DCAS-08656.) The JAQ asked the firefighters "if they performed each of 181 tasks, the importance of the task, the frequency of task performance, and whether the task would be performed on the first day of active duty if the occasion arose." (Id. at DCAS-08657.) The JAQ also asked firefighters to "rate the importance of 46 abilities and if they were required on the first day of the job." (Id.) Based on the JAQ results, Dr. Cline whittled the list down to 96 tasks and 40 abilities. (Id. at DCAS-08657-58.) Dr. Cline then grouped the 96 tasks into 16 task clusters. (Id.)

Dr. Cline convened a Linking Panel, consisting of 34 firefighters and fire officers, to assess the relationship between the identified tasks and abilities. (Id. at DCAS-08660.) Panelists used a five-point scale to rate the strength of the relationship between the task clusters and the different abilities. (Id.) Based on the results of this survey, Dr. Cline eliminated one ability, Written Expression. (Id.)

2. Test Specifications

Following the job analysis, a series of decisions were made – the record does not reveal by whom – concerning how Exam 6019 would be structured. (See Test Specifications Report (Def. Ex. A-4).) Among other things, it was decided that the exam would be a written exam only, and therefore would not test for abilities that could be measured only by using an oral or physical component; that the remaining 18 abilities and characteristics (collectively, "constructs") would be weighted equally; and that only candidates who passed the written exam would be permitted to take the physical exam. (Id. at DCAS-E-0386; Cline Rep. 8.)

3. Constructing the Exam

To create Exam 6019, Dr. Cline divided the 18 constructs into three groups: (1) cognitive abilities; (2) personal characteristics; and (3) abilities involving the speed at which candidates could process information ("timed abilities"). (See HT 48-50; Test Dev. Rep.) Each group was tested differently. Cognitive abilities were tested using standard multiple choice questions with clearly defined right or wrong answers. (See, e.g., Exam 6019 (Def. Ex. A-1) (filed under seal) at Exam 035.) Personal characteristics were tested using "situational judgment exercises" ("SJEs"), in which candidates were asked to rate how desirable it would be to take certain actions in response to certain situations. (Pl. PFF ¶ 90.) The timed abilities were tested using timed multiple-choice questions. (See Exam 6019 at Exam 002; HT 50.) Dr. Cline organized the constructs into groups as follows:

Cognitive Abilities (Cognitive Component)

- Problem Sensitivity ("the ability to tell when something is wrong or is likely to go wrong. It includes being able to identify the whole problem as well as elements of the problem.")
- Number Facility ("involves the degree to which adding, subtracting, multiplying and dividing can be done quickly and correctly. These can be steps in other operations like finding percentages and taking square roots.")
- Deductive Reasoning ("the ability to apply general rules to specific problems to come up with logical answers. It involves deciding if an answer makes sense.")
- Inductive Reasoning ("the ability to combine separate pieces of information, or specific answers to problems, to form general rules or conclusions. It involves the ability to think of possible reasons for why things go together.")
- Information Ordering ("the ability to follow correctly a rule or set of rules or actions in a certain order. The rule or set of rules used must be given. The things or actions to be put in order can include numbers, letters, words, pictures, procedures, sentences, and mathematical or logical operations.")
- Flexibility of Closure ("the ability to identify or detect a known pattern (like a figure, word, or object) that is hidden in other materials. The task is to pick out the disguised pattern from the background materials.")
- Spatial Orientation ("the ability to tell where you are in relation to the location of some object or to tell where the object is in relation to you.")
- Visualization ("the ability to imagine how something would look when it is moved around or when its parts are moved or rearranged. It requires the forming

of mental images of how patterns or objects would look after certain changes, such as unfolding or rotation. One has to predict how an object, set of objects, or pattern will appear after the changes have been carried out.")

Personal Abilities/Characteristics (SJE Component)

- Adaptability ("the ability to maintain effectiveness in varying environments, with various tasks, responsibilities, or people.")
- Tenacity ("the ability to stay with a position or plan of action until the desired objective is achieved or is no longer reasonably attainable.")
- Integrity ("involves the degree to which one can maintain social, ethical and organizations norms in job-related activities.")
- Work standards ("the ability to set high goals or standards of performance for self, subordinates, others and organization. Dissatisfied with average performance.")
- Resilience ("involves the degree to which one can handle disappointment and/or rejection while maintaining effectiveness.")
- Coordination ("the ability to adjust actions in relation to others' actions.")
- Establishing and Maintaining Interpersonal Relationships ("involves the degree to which one can develop constructive and cooperative w2orking relationships with others, and maintaining them over time.")

Speed Abilities (Timed Component)

- Speed of Closure ("involves the degree to which different pieces of information can be combined and organized into one meaningful pattern quickly. It is not known beforehand what the pattern will be. The material may be visual or auditory.")
- Perceptual Speed ("involves the degree to which one can compare letters, numbers, objects, pictures, or patterns, quickly and accurately. The things to be compared may be presented at the same time or one after the other. This ability also includes comparing a presented object with a remembered object.")
- Memorization ("the ability to remember information, such as words, numbers, pictures and procedures. Pieces of information can be remembered by themselves or with other pieces of information.")

(Job Analysis Rep. at DCAS-8687-6890.)

A panel of firefighters, working in conjunction with Dr. Cline and a DCAS Test and Measurement Specialist, drafted the test questions ("items" in testing parlance) for the cognitive component. (Test Dev. Rep. at DCAS 15230-31.) Dr. Cline, with the assistance of a DCAS employee, Barbara Kroos, Ph.D., developed the items for the SJE component. (Pl. PFF ¶ 86.) Dr. Bobko reviewed all of the items. (Test Dev. Rep. at DCAS 15232, 15234.)

4. Review and Pretesting

After the exam items had been drafted, Dr. Cline asked a group of 12 firefighters and lieutenants to review the items and complete rating sheets as part of a Final Subject Matter Expert ("SME") Review. The rating sheets asked the firefighters and lieutenants to indicate whether the items were appropriate for an entry-level examination and to rate the items on a five-point scale ranging from "Don't Use" to "Excellent." (Test Dev. Rep. at DCAS-15243.) The SME reviewers generally reacted negatively. Ten of the 12 firefighters and lieutenants rated all of the items in the SJE component (i.e., over 50% of the items on the exam) as "Post-Entry-Level" or "Don't Use." (Pl. Ex. 20; Pl. PFF ¶ 142.) Nine of the 12 reviewers wrote negative comments on the rating sheets, including the comments excerpted at the beginning of this opinion. (Pl. Ex. 20.) At the 6019 Hearing, Dr. Cline testified that she essentially ignored the SME Reviewers' negative ratings and comments because she did not believe that the firefighters were qualified to judge the utility of the SJE items. (HT 92-94.)

In late December 2006, around the same time as the Final SME Review, Dr. Cline conducted an "item classification" procedure with five DCAS Test and Measurement Specialists. (Test Dev. Rep. at DCAS-15234.) Dr. Cline gave the DCAS test specialists the list of constructs that Exam 6019 was supposed to measure, along with the definitions and examples Dr. Cline had used to define the constructs, and asked them to match each item on Exam 6019 to the construct it was supposed to measure. (Test Dev. Rep. at DCAS-15234; HT 68.) The purpose of the procedure was to determine whether the abilities and characteristics that Exam 6019 was intended to measure were sufficiently "operationalized" – that is, whether those constructs could be defined in terms of observable behaviors, as opposed to internal, mental processes that cannot be observed. (HT 68; Pl. PFF ¶ 52.) On the cognitive component of the exam, the test

specialists were able to correctly match the items to the intended constructs for only 62.5% of the 40 items. On the SJE component, the test specialists were able to correctly match only 60% of the 105 items to their intended constructs.[11] (Test Dev. Rep. at DCAS-15235-36.)

Finally, Dr. Cline administered a pretest to DCAS personnel on January 12, 2007. The limited purpose of the pretest was to determine whether the directions for the SJE and timed components were appropriate and how long the timed component should be. (Test Dev. Rep. 9.)

The City conducted these reviews only weeks before it administered Exam 6019. The DCAS test specialists attempted to classify the cognitive items in late December 2006, while Dr. Cline was on Christmas vacation. (HT 70.) The Final SME Review occurred on December 26 and 27, 2006. (Pl. Ex. 62.) According to an internal email, Exam 6019 was scheduled to go to print on December 28, 2006. (Pl. Ex. 14.) The test specialists attempted to classify the SJE items for the first time on January 10, 2007. (HT 77.) The City administered Exam 6019 ten days later on January 20, 2007.

B. Application of the Guardians Standard

This court analyzes the validity of Exam 6019 using the five-part test set forth in Guardians Association of the New York City Police Department, Inc. v. Civil Service Commission, 630 F.2d 79 (2d Cir. 1980). As the Second Circuit noted in 2006, Guardians is still the governing law in this Circuit, see Gulino v. New York State Educ. Dep't, 460 F.3d 361, 385 (2d Cir. 2006), cert. denied, 128 S. Ct. 2986 (2008), and the parties agree that it controls here.

As set forth more fully in the Disparate Impact Opinion, see 637 F. Supp. 2d at 109, Guardians represents both a distillation and a reconciliation of the differing test-validation approaches contained in the Equal Employment Opportunity Commission's "Uniform

[11] The test specialists did not review items on the timed component of the exam. (Test Dev. Rep. at DCAS-15235-36.)

Guidelines on Employee Selection Procedures" ("EEOC Guidelines"), 29 C.F.R. §§ 1607.1-1607.18. The EEOC Guidelines draw a sharp distinction between "tests that measure 'content' – i.e., the 'knowledges, skills or abilities' required by a job – and tests that purport to measure 'constructs' – i.e., 'inferences about mental processes or traits, such as 'intelligence, aptitude, personality, commonsense, judgment, leadership and spatial ability.'" Gulino, 460 F.3d at 384 (quoting Guardians, 630 F.2d at 91-92). A content validity approach to test validation measures the degree to which the content of a selection procedure is representative of important aspects of job performance. 29 C.F.R. § 1607.5(B). A construct validity approach assesses the degree to which the test measures identifiable characteristics that are important for successful performance of the job in question. Id. § 1607.14(D). The Guardians court observed, however, that content and construct exist on a single continuum that "starts with precise capacities and extends to increasingly abstract ones." Guardians, 630 F.2d at 93. The Guardians court therefore developed a "functional approach" to content validation that permits courts and employers to use a content validity approach to assess tests that measure some abstract abilities. Id. The court cautioned, however, that "the degree to which content validation must be demonstrated should increase as the abilities tested for become more abstract." Id.

Guardians lays out a five-part test for analyzing the content validity of an employment test:

(1) The test-makers must have conducted a suitable job analysis;
(2) They must have used reasonable competence in constructing the test itself;
(3) The content of the test must be related to the content of the job;
(4) The content of the test must be representative of the content of the job; and
(5) There must be a scoring system that usefully selects from among the applicants those who can better perform the job.

Id. The first two requirements relate to the "quality of the test's development," while the final three "are more in the nature of standards that the test, as produced and used, must be shown to have met." Id. The court turns now to the question of whether the City has carried its burden of demonstrating that it complied with these requirements.[12]

1. The City's Job Analysis

The EEOC Guidelines require employers who use a content validation approach to undertake a job analysis that assesses "the important work behavior(s) required for successful performance and their relative importance." See Guardians, 630 F.2d at 95 (quoting 29 C.F.R. § 1607.14C(2)). Here, the City's job analysis appears to meet that standard. Dr. Cline asked incumbent firefighters to generate a comprehensive list of 181 tasks and 46 abilities that she then used to construct a Job Analysis Questionnaire (JAQ) that she administered to over 340 firefighters. (HT 38-39; Job Analysis Report at DCAS-08652-56.) The JAQ asked respondents if they performed the listed tasks, how important the tasks were, how frequently they performed the tasks, and whether they would perform the tasks on the first day of active duty. (Id. at DCAS-08657.) It also asked the firefighters how important the listed abilities were and if they were required on the first day of active duty. (Id.) After using the JAQ results to narrow the task and abilities list, Dr. Cline convened a Linking Panel to rate the relationship between the tasks

[12] Although Plaintiffs argue that the City has failed to meet its burden under Guardians, they also argue in the alternative that this court should rule against the City without recourse to the Guardians factors. According to Plaintiffs, many of the constructs that Exam 6019 was meant to measure are internal, unobservable traits, which means that the exam cannot be properly assessed using content-validation techniques. Because the City relies exclusively on a content-validity approach to demonstrate that Exam 6019 is valid, Plaintiffs argue that it has necessarily failed to carry its burden of proof. (See generally Pl. Mem. 8-11.) While this argument has some force, the court considers it unwise to circumvent Guardians in this manner. First, Guardians and Gulino recognize that the line between observable abilities and internal traits can be hazy, and that courts should closely scrutinize an employer's actions before declaring that a particular validation method is categorically inappropriate. Second, the 18 constructs that Exam 6019 seeks to measure are not all abstract or "internal"; some, like Number Facility, are relatively concrete. In this sense, Exam 6019 exemplifies the continuum that Guardians sought to accommodate. Third, Plaintiffs' argument could just as easily apply to the question of whether the City has carried its burden at step three of the Guardians test to demonstrate that the content of Exam 6019 is related to the content of the job. See Guardians, 630 F.2d at 95. Accordingly, the court believes that the better course is to consider Plaintiffs' points about the relevance of the City's validation efforts as part of its Guardians analysis.

and abilities. She then used the Linking Panel results to eliminate one of the abilities from the list. (Id. at DCAS-08660.) This process substantially complies with both the EEOC and Guardians requirements. See 630 F.2d at 95-96.

In undertaking the Exam 6019 job analysis, the City appears to have avoided the two crucial errors that it made in its job analysis for Exams 7029 and 2043.[13] As detailed in the Disparate Impact Opinion, the City's previous job analysis (1) failed to demonstrate any relationship between abilities and tasks, and (2) retained tasks and abilities that could be learned on the job, as opposed to tasks and abilities that firefighters must perform or possess prior to appointment. See D.I. Op., 637 F. Supp. 2d at 108-113. Here, the Linking Panel, under the supervision of Dr. Cline, appears to have understood the definitions and importance of the various abilities and meaningfully analyzed the abilities' relationship to the listed tasks. (See Job Analysis Report at DCAS-08660 and Appx. B.) By seeking information about "Day One" tasks and abilities on the JAQ, the City also appears to have made reasonable efforts to eliminate tasks and abilities that could be learned on the job.

The City also did a better job of selecting relevant abilities and characteristics for testing. As described in the Disparate Impact Opinion, the designers of Exams 7029 and 2043 sought to test only nine abilities, all of which were cognitive. In doing so, the City ignored a host of non-cognitive abilities that are equally, if not more important to the job. See D.I. Op., 637 F. Supp. 2d at 119-121. By contrast, Dr. Cline consulted numerous reputable sources to compile a comprehensive list of abilities, (HT 36-37), and used a reasonable review process to narrow the list to 18 non-physical constructs. In doing so, Dr. Cline retained many of the constructs that this

[13] The City used the same job analysis for Exam 7029 and 2043.

court previously criticized the City for excluding.[14] (See Job Analysis Report; D.I. Op., 637 F. Supp. 2d at 119.)

2. The Test Construction Process

The second Guardians requirement is that the test-makers must use reasonable competence in constructing the test. As described in the Disparate Impact Opinion, Guardians offers two points of guidance on this issue. First, civil service examinations should be constructed by testing professionals rather than lay employees. Guardians, 630 F.2d at 96 ("[E]mployment testing is a task of sufficient difficulty to suggest that an employer dispenses with expert assistance at his peril."). Second, exam questions should be tested on a sample population to roughly determine whether they accurately measure the targeted abilities and to ensure that they are comprehensible and unambiguous. Id.

The City retained a test-preparation specialist, Dr. Cline, to oversee the construction process. But just as it did with Exams 7029 and 2043, the City erred by permitting firefighters to write test items, in this case for the cognitive component of the exam. (See Pl. PFF ¶¶ 75-77; D.I. Op., 637 F. Supp. 2d at 115.) Guardians explicitly warns employers not to leave item-drafting to lay employees, "who may have . . . expertise in identifying tasks involved in their job but [are] amateurs in the art of test construction." 630 F.2d at 96. It is no defense to say, as the City does, that the items written by firefighters were later reviewed by an expert, since this was precisely the process condemned in Guardians. Id. at 84, 96.

The City also sample-tested Exam 6019 on firefighters during the Final SME Review. That review occurred so late in the construction process, however, that the City was unable to

[14] Plaintiffs assert that the City failed to test for an important ability, Mechanical Comprehension. (See, e.g., Pl. PFF ¶ 70.) The City is not required to test every single potentially useful ability. Guardians, 630 F.2d at 99. Plaintiffs may address the utility and feasibility of testing for Mechanical Comprehension with the Special Master when developing the future exam.

integrate its findings into a revised exam. Dr. Cline conducted the Final SME Review on December 26 and 27, 2006 – one day before Exam 6019 was scheduled to go to print.[15] (Pl. Exs. 14, 62; HT 82.) As described above, the SME reviewers provided a large amount of negative feedback that, had it been actually considered by the test designers, would have required a substantial rethinking or reorganization of the exam. Dr. Cline admitted at the 6019 Hearing, however, that she largely ignored the firefighters' ratings and comments, and that at any rate it was "way too late" to alter the major components of the exam. (HT 92-95; see also fn. 14, above.)

Dr. Cline also administered a pretest to DCAS personnel on January 12, 2007 to test the exam instructions and the length of the timed section – but not to test whether the questions were comprehensible or accurately measured the 18 constructs. (Test Dev. Rep. 9.) For Guardians purposes, this pretest suffers the triple infirmity of having been given too late, for the wrong reasons, and to the wrong people. Therefore, despite its belated efforts, the City failed to comply with Guardians' requirement to conduct appropriate sample testing.

3. Relationship Between Test Content and Job Content

Guardians' third requirement that the content of the exam be directly related to the content of the job reflects "[t]he central requirement of Title VII" that a test be job-related. 630 F.2d at 97-98. The correspondence between exam and job is a function of two conditions: whether the constructs that the employer sought to test were required for successful job performance, and whether the exam actually tested those abilities. See D.I. Op., 637 F. Supp. 2d at 117. As detailed above in Section IV.B.1, the City has successfully demonstrated that the 18 constructs it sought to test were related to the job of entry-level firefighter. But the City has

[15] The City does not dispute that Exam 6019 went to print on December 28, 2006. Although Dr. Cline testified at the 6019 Hearing that she made changes to exam items on the basis of the Final SME Review results, she also admitted that she did not know when the exam was printed. (HT 204.)

failed to demonstrate that Exam 6019 actually tested those 18 constructs. This failure is fatal to the City's showing under Guardians' third requirement and, ultimately, to its ability to demonstrate that Exam 6019 is valid.

i. Appropriateness of a Content-Validity Approach

The EEOC Guidelines, and test-construction professionals generally, recognize at least three distinct methods of test validation: criterion-related validation, construct validation, and content validation. (See Pl. Ex 18 ¶ 49; HT 32.) Criterion-related validation examines whether candidate performance on an exam correlates with actual job performance. (Pl. Ex 18 ¶ 51; HT 33.) This inquiry is retrospective, and requires job-performance data such as appraisals or supervisor ratings. (Pl. Ex 18 ¶ 51.) Construct validation evaluates an exam's relationship to other well-researched or validated exams, including whether the exams share common characteristics and whether performance on a given exam correlates with performance on another exam that has been shown to have criterion-related validity in a similar setting. (Pl. Ex 18 ¶¶ 58-59; HT 33.) Content validation, according to Plaintiffs' expert, "seeks to use logical, and sometimes statistical, evidence to evaluate whether the selection procedure can be viewed as 'mapping onto' and calling for a candidate to display behaviors that are important to performing the job." (Pl. Ex 18 ¶ 53; see also HT 33 (testimony of Dr. Cline that content validity "involves looking at the test content and see[ing] whether it matches something that it's desirable to match").)

The City relied solely on a content-validity approach to validate Exam 6019. (Id. 34; Pl. PFF ¶ 48.) Dr. Cline conducted a job analysis, identified relevant constructs, and attempted to create items that measured those constructs – all hallmarks of a content-validation approach.

(See, e.g., HT 41-43.) Plaintiffs argue that this approach was fundamentally inappropriate, and that as a result the City cannot prove that Exam 6019 is valid.

According to Plaintiffs' expert, Dr. Jones, "[c]ontent validation becomes an increasingly appropriate strategy as the opportunity to show a linkage between observable aspects of the selection procedure and observable aspects of the job increases. The ubiquitous example is content validation of a computer-based word processing test to assess candidates where word processing is a key part of the job."[16] (Pl. Ex 18 ¶ 54.) Conversely, Guardians, the EEOC Guidelines, and Dr. Jones's testimony all make clear that content validation is an inappropriate validation strategy if the exam seeks to measure internal, unobservable mental traits or constructs. See Guardians, 630 F.2d at 93, 98; Pl. Ex 17 ¶¶ 39-40; HT 308-311. The EEOC Guidelines state that:

> A selection procedure based upon inferences about mental processes cannot be supported solely or primarily on the basis of content validity. Thus, a content strategy is not appropriate for demonstrating the validity of selection procedures which purport to measure traits or constructs, such as intelligence, aptitude, personality, commonsense, judgment, leadership, and spatial ability.

29 CFR § 1607.14C(1). This prohibition reflects the considered judgment of courts, agencies, and test-construction experts that it is extremely difficult to determine ex ante whether a multiple-choice, paper-and-pencil exam accurately measures internal mental processes or personal attributes.[17] Thus, according to Dr. Jones, test-makers seeking to measure such abstract constructs typically use a criterion-related or construct validation approach in place of, or in conjunction with, a content-based approach. (Pl. Ex. 17 ¶ 48.)

[16] Dr. Cline offered a similar example at the 6019 Hearing: "For example, if a test for algebra matches the content of an algebra course. That would be content validity." (HT 32.)

[17] This is not to say that an exam can never measure such traits – only that it is nearly impossible to tell if it will do so before it has been administered.

Here, it is clear that Exam 6019 seeks to measure abstract, unobservable mental constructs such as Flexibility of Closure, Speed of Closure, and Problem Sensitivity, as well as personality characteristics like Integrity, Adaptability, Tenacity, Work Standards, and Resilience. These are precisely the sort of "traits or constructs" that render an exam unfit for content-based validation. The City's exclusive reliance on a content validation approach necessarily means that it has failed to demonstrate that the content of Exam 6019 is related to the content of the job of entry-level firefighter.

The City argues that a content-validation approach is appropriate in this case because Dr. Cline sufficiently "operationalized" the traits and mental constructs she sought to measure – i.e., she defined them in terms of observable behaviors rather than unobservable internal processes. (HT 67.) This argument is undercut by the results of the item classification procedure, in which Dr. Cline asked five DCAS Test and Measurement Specialists to match each item on Exam 6019 to the ability it measured. As Dr. Cline admits, the entire purpose of the classification procedure was to determine whether the Exam 6019 constructs were sufficiently operationalized. (HT 68.) The DCAS specialists, however, were unable to correctly classify approximately 40% of the items. (Test Dev. Rep. at DCAS-15235-36.) This result should not have been surprising: as described above, the inherent difficulty of ascertaining what an exam measures before it is administered is precisely what animates the Guidelines' prohibition against using content validation to support exams that measure internal traits or unobservable mental constructs. In this case, the City's own data demonstrate the inappropriateness of relying on a content-validity approach to assess Exam 6019, above and beyond the admonitions in the EEOC Guidelines and professional literature.

ii. Factor and Reliability Analyses

After the candidates took Exam 6019, Dr. Cline conducted a "factor analysis" and a "reliability analysis" on the scoring data to determine whether the exam measured the constructs it was intended to measure. (See Exam Analysis.) According to the City, the results of these analyses demonstrate that Exam 6019 accurately measured whether candidates possessed the 18 abilities and characteristics that Dr. Cline identified. Plaintiffs' evidence demonstrates the opposite.

According to Dr. Jones, factor analysis is a means of identifying patterns in a set of data – in this case, the candidates' answers to each item on Exam 6019.[18] (HT 323.) Factor analysis is used to determine whether exam items group together into "factors" or clusters consisting of items intended to measure the same construct. The underlying assumption is that a given candidate is likely to perform consistently on similar items; for example, if a candidate correctly answers a Spatial Reasoning question, that candidate is more likely to correctly answer other questions that measure Spatial Reasoning. The degree to which candidates' answers aggregate into identifiable clusters, and the number of such clusters, can be compared to the employer's exam plan to determine whether the exam actually measured what it was supposed to measure. (Id.; see also D.I. Op., 637 F. Supp. 2d at 116.) For example, had the City designed Exam 6019 perfectly, a factor analysis of the 195 items on the exam would reveal 18 clusters, each defined by the items written to assess one of the 18 constructs. (See, e.g., Pl. Ex. 17 ¶¶ 140-41.)

Although Dr. Cline conducted a factor analysis, she did not factor analyze the individual items on Exam 6019. Instead, she sorted the 195 items into 18 subgroups, based on which construct each item was intended to measure, and added up candidates' scores on each subgroup

[18] The City's submissions do not explain the purpose or methodology of factor analysis.

of items to create 18 "subscores." Dr. Cline then factor analyzed those subscores. (Id. ¶ 135; HT 99-100, 330.) As Plaintiffs point out, by factor analyzing the subscores rather than the individual items, Dr. Cline's factor analysis assumed, rather than proved, that the sets of items written to measure each ability correlated with one another more strongly than they correlated with other sets of items that were written to assess other constructs. (Pl. PFF ¶ 124 (emphasis in original).) In other words, Dr. Cline's factor analysis did not do what the City needed it to do: reveal approximately 18 clusters, each composed primarily of items written to test identical constructs. What Dr. Cline's subscore factor analysis did do, according to her testimony, is reveal a strong correlation among items within each of the three broad components of the exam – cognitive, SJE, and timed. (HT 102-04.) This result is irrelevant. The question is whether Exam 6019 accurately measured the 18 constructs that it was designed to measure, not the three external categories that those constructs were grouped into.

Dr. Jones, by contrast, conducted an "item-level" factor analysis of Exam 6019. (Pl. Ex. 17 ¶ 138.) His factor analysis found that items that were intended to measure completely different constructs clustered together, and, conversely, that items that were intended to measure identical constructs were scattered across clusters rather than bunched together. (See id. ¶¶ 145-46, 220-21.) Dr. Cline disputes the accuracy of his findings, but even if her contention were correct, it would not alter the fact that her factor analysis does not show that Exam 6019 did what it was supposed to do. The burden to make this showing lies with the City.

Dr. Cline also conducted a reliability analysis on the Exam 6019 scoring data. (See Exam Analysis.) According to Dr. Jones' testimony, reliability analysis, also called internal consistency analysis, is in some senses a mirror version of factor analysis. Instead of looking for patterns in the scoring data as a whole, reliability analysis pulls out sets of items that were meant

to correspond to one another according to the test plan – in this case, the 18 subgroups of items that corresponded to each intended construct. (HT 333.) The reviewer then assesses the degree to which the test-takers performed consistently on the items in each subgroup. (Id.)

Testing professionals use a statistic called "coefficient alpha" to calculate how reliable each subgroup is. (Pl. Ex. 17 ¶ 158.) According to Dr. Cline, coefficient alpha measures how much error there is in a test score. (HT 111.) A coefficient alpha of .60 indicates that 40% of the score is due to error variance. (Id.) A coefficient alpha of .70 or greater is typically accepted as indicating a reliable measure of a given construct. (Pl. Ex. 17 ¶¶ 158-59.) When Dr. Cline calculated the internal consistency of the 18 subgroups of Exam 6019 items, she obtained the following results:

Cognitive Ability the Items Were Intended to Measure	**No. of Test Items**	**Coefficient Alpha Reliability**
Deductive Reasoning	4	.36
Flexibility of Closure	5	.43
Information Ordering	5	.57
Inductive Reasoning	5	.45
Number Facility	5	.50
Problem Sensitivity	6	.42
Spatial Orientation	5	.48
Visualization	5	.48
Speed of Closure	5	.70
Perceptual Speed	40	.95
Memory	5	.26
Adaptability	15	.44
Coordination	15	.54
Establishing and Maintaining Interpersonal Relationships	15	.40

(Exam Analysis at DCAS-19483-19484.) Only two subgroups of items – those intended to measure Speed of Closure and Perceptual Speed – had coefficient alphas of .70 or greater.

Twelve of the 18 subgroups had coefficient alphas of .50 or lower, indicating that 50% or more of the score was due to error. The remaining four subgroups had coefficient alphas between .53 and .57, indicating that 43% to 47% of the score was due to error. According to Dr. Jones, these reliabilities are no better than those produced when he randomly selected items to be placed in each subgroup using a random number generator. (Id. ¶¶ 168-71.)

Dr. Cline asserted at the 6019 Hearing that these calculations do not indicate that Exam 6019 failed to test for the intended constructs, because "when the [subscores] are correlated all together and added into a single score, the overall score is going to be reliable." (HT 112; see also Cline Expert Rep. 9 ("The reliability of the total score was .94, which . . . is generally acceptable.").) Dr. Cline did not explain or defend this claim, which is problematic for the City, given that it bears the burden of demonstrating job-relatedness. What is clear is that Dr. Cline admitted on the stand that her reliability analysis does not say "anything" about whether Exam 6019 measures 18 specific constructs. (HT 113.) In this phase of the Title VII inquiry, that concession is dispositive of the issue.

In sum, the City has not established that Exam 6019 did what it was designed to do: determine whether candidates possessed the 18 abilities and characteristics that the City deemed necessary to the job of entry-level firefighter. As confirmed by the classification procedure, the City's reliance on a content-based approach to validate an exam that tested internal, unobservable constructs was fundamentally misplaced. To the extent that content validation was appropriate for the portions of the exam that tested observable constructs, the City's statistical analyses undermine any contention that the exam actually measured those constructs, and Dr. Jones's statistical analyses refute it.

4. Representativeness of the Test Content

The fourth Guardians requirement is that a test be a "representative sample of the content of the job." 630 F.2d at 98 (internal quotation marks omitted). This requirement has two components: "[t]he first is that the content of the test must be representative of the content of the job; the second is that the procedure, or methodology, of the test must be similar to the procedures required by the job itself." Id. These requirements are not strict, since rigorous interpretation of either component would "foreclose any possibility of constructing a valid test." Id.

With respect to content representativeness, "it is reasonable to insist that the test measure important aspects of the job, at least those for which appropriate measurement is feasible, but not that it measure all aspects, regardless of significance, in their exact proportions." Id. at 99. The reason for this requirement is to "prevent either the use of some minor aspect of the job as the basis for the selection procedure or the needless elimination of some significant part of the job's requirements from the selection process entirely; this adds a quantitative element to the qualitative requirement that the content of test be related to the content of the job." Id.

Plaintiffs argue, and this court agrees, that the City's complete failure to carry its burden at step three of the Guardians analysis necessarily means that it has not demonstrated content representativeness at step four. As detailed above, the City cannot establish that Exam 6019 accurately measures any of the constructs that it was intended to measure. A fortiori, the City cannot establish that the exam measures any "important aspects of the job." This does not mean that the fourth Guardians requirement is redundant: had the City demonstrated that Exam 6019 tested some or all of the intended constructs, this court would be required to undertake the additional "quantitative" inquiry into whether "some minor aspect of the job" was used as a basis

for selection or "some significant part of the job's requirements" was eliminated from the exam. See id. But where the employer has failed to show any significant correlation between exam content and job content, the content-representativeness component of step four is extraneous.

The requirement that a test's procedures be representative is intended "to prevent distorting effects that go beyond the inherent distortions present in any measuring instrument." Id. In particular, "although all pencil and paper tests are dependent on reading, even if many aspects of the job are not, the reading level of the test should not be pointlessly high." Id. Dr. Cline analyzed the reading level of Exam 6019 after this court's July 2009 liability ruling, which criticized the unnecessarily high reading level of Exams 7029 and 2043. (See Readability Rep. (Def. Ex. A-7); D.I. Op., 637 F. Supp. 2d at 122-23). Dr. Cline compared Exam 6019 with the Probationary Firefighter's Manual that the FDNY uses at its Fire Academy. (Readability Rep.) According to Dr. Cline, the median reading grade level for Exam 6019 is 7.7, while the median reading level of the Manual is 10.0. (Id.) This analysis sufficiently demonstrates that the reading level of Exam 6019 is not too high. See D.I. Op., 637 F. Supp. 2d at 123 (reading level of written test questions "should be below the reading level of material Firefighters may be given to read at the Fire Academy").

5. Utility of the Scoring and Selection System

Under the fifth Guardians requirement, the City must demonstrate that its scoring system usefully selects those candidates who can better perform the job of entry-level firefighter. Guardians, 630 F.2d at 100. The court notes at the outset that it is approaching this analysis in a different posture than the Guardians court. In Guardians, the Second Circuit held that the defendant had submitted adequate evidence of content-relatedness and representativeness under requirements three and four – in other words, that the challenged exam accurately assessed

important job constructs. 630 F.2d at 99. The Guardians court therefore went on to consider whether the defendant's use of a cutoff score and a rank-order mechanism was justified by business necessity. Id. at 100, 105. Here, by contrast, the City has failed to make its showing under requirements three and four. The City has not established that Exam 6019 reliably measures any of the abilities or characteristics required by entry-level firefighters. This fact obviates the need to inquire into the City's scoring system. If candidates' exam scores provide no useful information about their qualifications, then by definition there is no "valid" cutoff score that will usefully differentiate between qualified and unqualified candidates, and no "valid" justification for ranking candidates based on their exam scores. Step five of Guardians is only appropriate for determining whether an otherwise valid exam was used improperly. See Barbara Lindemann & Paul Grossman, Employment Discrimination Law 158 (3d ed. 1996). That situation is not present here.

Even if the City could demonstrate that Exam 6019 has some level of validity, it could not establish that its use of the exam as a pass/fail and rank-ordering device is job-related. With respect to pass/fail use, the City admits that it used a cutoff score of 70 on Exam 6019 because 70 is the default cutoff score provided by the City's civil service rules. (Pl. PFF ¶ 12.) As this court held in the Disparate Impact Opinion, the City may not simply rely on civil service rules to set a passing score. Instead, the City must present evidence that its chosen cutoff score bears a direct relationship to the necessary qualifications for the job of entry-level firefighter. D.I. Op., 637 F. Supp. 2d at 125. The City has not attempted to make such a showing.

The City, to its credit, acknowledges that its chosen cutoff score is improper. (See Def. Mem. 12.) The City suggests, instead, that it would be appropriate to use an entirely different cutoff score to distinguish qualified from unqualified candidates. At the 6019 Hearing, Dr. Cline

testified that she reanalyzed the Exam 6019 data using something called a "Modified Angoff Method," and that the results of this reanalysis show that a cutoff score of 80, rather than 70, is appropriate – i.e., job-related. (HT 55-57; see also Angoff Report.) The City now contends that using Exam 6019 on a pass/fail basis with a cutoff score of 80 is a valid selection procedure under Title VII.

The court is perplexed by the City's approach. The City is essentially asking the court to ratify a hypothetical use of Exam 6019, rather than the City's actual use of the exam. But the City could have done any number of things with the exam results. It could have banded candidates' scores, or combined them with candidates' CPAT scores, or decided that there was a strong basis in evidence that Exam 6019 violates Title VII and then discarded the exam scores altogether. The City took none of these actions, so the court does not need to consider them. And for the same reason, the court does not need to consider what would have happened if the City had set the cutoff score at 80. The only question before the court is whether what the City actually did – use Exam 6019 as a pass/fail device with a cutoff score of 70 – was justified by business necessity. If, as the City admits, that action was not justified – if the cutoff score that it used did not differentiate between qualified and unqualified candidates – then the inquiry is over. The candidates who were harmed by the City's invalid cutoff score have already been harmed, and the candidates who benefited from the invalid cutoff have already benefited. The only question now is what the court should do to fix this problem.

Viewed in proper perspective, the City's proposal to use Exam 6019 with a different cutoff score is actually a proposed remedy for its admittedly invalid selection procedure. The practical effect of the City's proposal is that 2,100 candidates who scored between 70 and 80 on Exam 6019 would be retroactively stricken from the eligibility list. (See Def. Mem. 13.) As the

City admits, implementing the new cutoff score would mean that "about 15% of the candidate pool would fail, with the result being a greater impact on minorities." Id. The benefit, according to the City, is that the new pass/fail system would be job-related. But for the reasons outlined above, the City cannot demonstrate that any cutoff score, however derived, would be valid. The problem lies with the exam, not the cutoff. The court therefore declines to implement the City's proposed remedy.

With respect to rank-ordering, Dr. Cline asserts that her Angoff analysis demonstrates that "[d]ifferences in test scores . . . differentiate different types of candidates and provide support under the [EEOC] Guidelines for the use of Examination 6019 in a rank-order fashion." (Angoff Report at DCAS 19537.) Mere differentiation, however, is insufficient to support the rank-ordering of an eligibility list containing 21,000 candidates, the vast majority of whom are bunched tightly at the top.[19] "Permissible use of rank-ordering requires a demonstration of such substantial test validity that it is reasonable to expect one- or two-point differences in scores to reflect differences in job performance. . . . Without some substantial demonstration of reliability it is wholly unwarranted to make hiring decisions, with a disparate racial impact, for thousands of applicants that turn on one-point distinctions among their passing grades." Guardians, 630 F.2d at 100-101. Dr. Cline admitted in a deposition, however, that her Angoff analysis provides no information about the expected difference in job performance associated with even a three-point difference in scores on Exam 6019. (Pl. PFF ¶ 268.) Nor could it, given that the City has failed to demonstrate that Exam 6019 is content valid.

[19] According to Plaintiffs' analysis, of the top 10,000 candidates on the eligibility list, only five scored below 80 on the exam. (Pl. Ex. 4a ¶ 15.) The 2,039th ranked candidate on the eligibility list has an Adjusted Final Average score of 98.727. (Id. ¶ 18.) The 886th ranked candidate has an Adjusted Final Average score of 100.593. (Id. ¶ 16.) In other words, fewer than two points separate the 1,153 highest-ranked remaining candidates.

V. CONCLUSION

The parties do not dispute that the City's use of Exam 6019 to select entry-level firefighters disproportionately impacts black and Hispanic firefighters. For the reasons stated above, the City has not met four of the five requirements that it must meet to demonstrate that Exam 6019 is content-valid under Guardians Assoc. of New York City Police Dept., Inc. v. Civil Service Comm'n., 630 F.2d 79, 100 (2d Cir. 1980). The City has therefore failed to carry its burden to demonstrate that its use of Exam 6019 as a pass/fail and rank-ordering device is job-related and justified by business necessity. The court therefore concludes that the City's use of Exam 6019 does not comply with Title VII.

Accordingly, the court restrains and enjoins the City from taking any further steps to initiate or finalize a fire academy class using the Exam 6019 eligibility list until October 1, 2010. The court will hold a hearing at the earliest possible date to consider the remedial measures it should take in light of this order.

SO ORDERED.

/s/ Nicholas G. Garaufis
NICHOLAS G. GARAUFIS
United States District Judge

Dated: Brooklyn, New York
August 4, 2010